STUDIA UBB CHEMIA, LIX, 1, 2014 (p. 99-110)
(RECOMMENDED CITATION)

QSAR STUDY ON NITROGEN-CONTAINING FLAVONOIDS BY SIMILARITY CLUSTER PREDICTION

TEODORA ELENA HARSA[a]

ABSTRACT. A set of 40 flavonoids, downloaded from the PubChem database, was submitted to a qsar study by using an alignment procedure of the molecules over the hypermolecule, that mimics the investigated correlational space, within the correlation weighting analysis. The best models have been validated in the external test set and in a new version of validation/prediction by using similarity clusters.

Key-words: QSAR, log P, flavonoids, correlation weighting, similarity.

1. INTRODUCTION

Flavonoids are phenolic substances with a low molecular weight and they are abundant in plant tissues, fruits (particularly in the skin) [1,2]. In the human body they manifest a lot of biological properties, such as antioxidants, antiallergenic, antibacterial, antifungal, antiviral and anticarcinogenic agents. These characteristics confer to flavonoids pharmacological properties useful in the treatment of diseases, ranging from allergies, bacterial and viral infectious processes to those of greater risk like the coronary diseases, cancer and HIV [3-4].

In the past half century, the use of QSAR (quantitative-structure-activity-relationship, one of the well-developed areas in computational chemistry) [5] has become increasingly helpful in understanding many aspects of chemical–biological interactions in drug and pesticide research, particularly enzyme functions, as well as in the areas of toxicology [6-8].

The parameter correlated in this paper is logP, the (calculated) partition coefficient in octanol/water, a measure of hydrophobicity that gives information about the transport of a drug through the cell membranes to the biological receptor [9].

[a] *Babes-Bolyai University, Department of Chemistry Faculty of Chemistry and Chemical Engineering, Arany Janos 11, 400028, Cluj, Romania, harsa.teodoraelena@yahoo.com*

Any similarity measure has three principal components: (i) the representation that is used to describe each of the structures taken in the work; (ii) the weighting scheme, used to assign weights to different parts of the structure representation that reflect their relative degrees of importance; and (iii) the similarity coefficient, used to quantify the degree of resemblance between two suitably weighted representations [10].There is a variety of ways for computing the similarity score;it is recognized that there is no a single similarity measure that will provide optimal screening in all circumstances [11−13].

2D similarity approaches can be defined as simple methods since they employ topological measurements derived from molecular graphs, but many times they show inconsistencies for the appropriate representation of QSAR/ QSPR predictive spaces [14].

Correlation weighting was previously discussed by Toropov and Toropova [15,16]

Graph theoretical descriptors, invariants up to isomorphism, also called topological indices [17-19] are used as predictor variables in qsar studies. Within this paper we used the indices generated by the TopoCluj software [20].

2. DATA SET

A set of 40 molecular structures, belonging to the class of flavonoids, have been downloaded from the database Pubchem [21] (Table 1), together with their log P. The set was split in the training set and test set (25 and 15 molecules, respectively). The structures have been optimized by HyperChem software, at molecular mechanics MM+ and semi-empirical PM3 levels of theory.

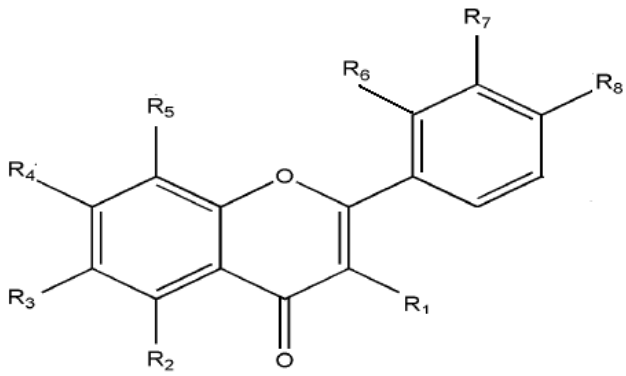

Table 1. The set of flavonoids, taken from PubChem database [21].

	Substituents	**log P**		**Substituents**	**log P**
1	R_2-NH_2	3.4	2	R_1-H_2, R_3-OH, R_8-NH_2	2.9
3	R_1-H_2, R_8-NH_2	2.5	4	R_2-NH_2, R_6-NH_2	2.4
5	R_2-NH_2, R_8-$N(Me)_2$	2.2	6	R_2-NH_2, R_8-$N(Et)_2$	3
7	R_3-NH_2	3.3	8	R_2-NH_2, R_8-NH(Et)	2.5
9	R_4-NH_2	3.3	10	R_1-NH_2	3.1
11	R_1-NH_2, R_3-NH_2	2.4	12	R_2-OH, R_5-NH_2	2.7
13	R_3-NH_2, R_4-OH, R_5-NH_2	1.5	14	R_1-NH_2, R_6-NH_2	2.4
15	R_6-NH_2, R_7-Me	3.2	16	R_2-OH, R_4-OH, R_7-NH_2 , R_7-OMe	1.4
17	R_2-OH, R_3-NH_2, R_4-OH	1.4	18	R_2-OH, R_5-NH_2, R_8-NH_2	1.1
19	R_2-OH, R_4-OH, R_7-NH_2 , R_8-OH	1.1	20	R_4-OH, R_5-NH_2 , R_8-NH_2	1.8
21	R_3-OH, R_8-NH_2	2.9	22	R_3-NH_2, R_8-$N(Me)_2$	3
23	R_3-OMe, R_8-NH_2	2.8	24	R_7-OMe, R_8-NH_2	2.8
25	R_3-Me, R_4-Me, R_8-$N(Me)_2$	4.4	26	R_2-NH_2, R_8-OH	1.8
27	R_7-NH_2, R_8-NH_2	2.2	28	R_2-NH_2, R_8-NH(Me)	2.1
29	R_2-NH_2, R_6-OH, R_8-OH	2.3	30	R_3-Me, R_8-$N(Me)_2$	4
31	R_2-NH_2, R_6-$N(Me)_2$	3.2	32	R_2-NH_2, R_6-NH(Me)	3
33	R_2-NH_2, R_6-OH	2.7	34	R_2-NH_2, R_6-OMe	3
35	R_2-NH_2, R_5-OH, R_6-NH_2	2	36	R_2-NH(Me), R_6-$N(Me)_2$	3.8
37	R_2-NH(Me), R_8-$N(Me)_2$	2.9	38	R_2-NH_2, R_8-NH_2	1.4
39	R_2-NH_2, R_7-OH	2.7	40	R_2-NH_2, R_6-OH, R_8-NH_2	2

3. METHOD

In performing the QSAR, we followed an algorithm based on the alignment of molecules over a hypermolecule [22] and correlation weighting analysis [15,16]. This algorithm includes the main steps: (a) download from PubChem (or other public domain) a dataset of molecules and optimize them at a choice level of theory; (b) compute global and local quantum and/or topological descriptors; (c) build up the hypermolecule of the set by superimposing the common and distinct features of all the molecules; (d) split the set of molecules in the learning and test sets, respectively; (e) write the binary vecteors, with 1 if there exists a fragment in the current molecule in a given position of the hypermolecule and 0 otherwise; (f) weight the binary vectors by various physico-chemical or mathematical local properties; (g) data reduction: discard the non-variant descriptors and (statistically) non-significant descriptors X_j over the j^{th} position of the hypermolecule; (h) make correlation weighting (including all significant positions j of the hypermolecule) and generate correlation weights CD_{ij} (as product of local descriptors – e.g. charge, mass, etc. with the regression coefficients for the significant positions of the hypermolecule), next sum them to give a global descriptor $SD_i = \sum_j CD_{ij}$; (i) models generation (i.e. QSPR/QSAR equations) by using various global descriptors; (j) validation of the model, either by the leave-one-out LOO (or similar procedures) or by the (external) test set; (k) validation by clusters of similarity.

3.1. Alignment over the hypermolecule

On the set of 40 flavonoids, a Hypermolecule [22] was built up, as a union of the common and distinct substructures over all the molecules in the set (Figure 1). Molecules were aligned over the hypermolecule positions and binary vectors were constructed (see Table 2, somesignificant positions), with 1 when in the current molecule there exists a corresponding atom and zero, otherwise.Next, the values 1 are replaced with local characteristics: partial charges, mass fragments or local topological descriptors. We used here the mass fragments in building the weighted vector for every molecule.

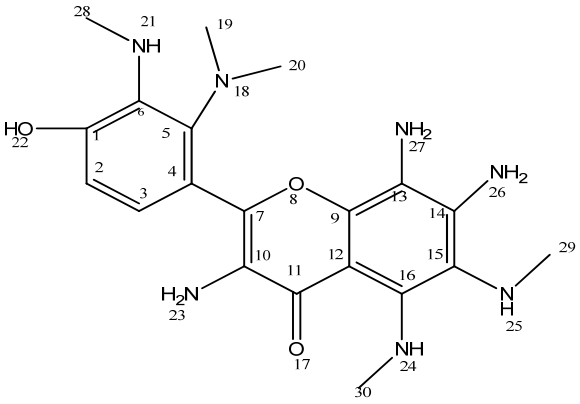

Table 2. Binary vectors cf. the hypermolecule, for the 40 flavonoids.

Molecule	17	18	19	20	21	22	23	24	25	26	27	28	29	30
1	1	0	0	0	0	0	0	1	0	0	0	0	0	0
2	1	0	0	0	0	0	0	0	0	0	0	0	0	0
3	1	0	0	0	0	0	0	0	0	0	0	0	0	0
4	1	1	0	0	0	0	0	1	0	0	0	0	0	0
5	1	0	0	0	0	0	0	1	0	0	0	0	0	0
6	1	0	0	0	0	0	0	1	0	0	0	0	0	0
7	1	0	0	0	0	0	0	0	1	0	0	0	0	0
8	1	0	0	0	0	0	0	1	0	0	0	0	0	0
9	1	0	0	0	0	0	0	0	0	1	0	0	0	0
10	1	0	0	0	0	0	1	0	0	0	0	0	0	0
11	1	0	0	0	0	0	1	0	1	0	0	0	0	0
12	1	0	0	0	0	0	0	0	0	0	1	0	0	0
13	1	0	0	0	0	0	0	0	1	0	1	0	0	0
14	1	0	0	0	0	0	1	0	0	0	0	0	0	0
15	1	1	0	0	0	0	0	0	0	0	0	0	0	0
16	1	0	0	0	0	1	0	0	0	0	0	0	0	0
17	1	0	0	0	0	0	0	0	1	0	0	0	0	0
18	1	0	0	0	0	0	0	0	0	0	1	0	0	0

Molecule	17	18	19	20	21	22	23	24	25	26	27	28	29	30
19	1	0	0	0	1	1	0	0	0	0	0	0	0	0
20	1	0	0	0	0	0	0	0	0	0	1	0	0	0
21	1	0	0	0	0	0	0	0	0	0	0	0	0	0
22	1	0	0	0	0	0	0	0	1	0	0	0	0	0
23	1	0	0	0	0	0	0	0	0	0	0	0	1	0
24	1	0	0	0	0	0	0	0	0	0	0	1	0	0
25	1	0	0	0	0	0	0	0	0	0	0	0	0	0
26	1	0	0	0	0	1	0	1	0	0	0	0	0	0
27	1	0	0	0	1	0	0	0	0	0	0	0	0	0
28	1	0	0	0	0	0	0	1	0	0	0	0	0	0
29	1	0	0	0	0	1	0	1	0	0	0	0	0	0
30	1	0	0	0	0	0	0	0	0	0	0	0	0	0
31	1	1	1	1	0	0	0	1	0	0	0	0	0	0
32	1	1	1	0	0	0	0	1	0	0	0	0	0	0
33	1	0	0	0	0	0	0	1	0	0	0	0	0	0
34	1	0	1	0	0	0	0	1	0	0	0	0	0	0
35	1	1	0	0	0	0	0	1	0	0	0	0	0	0
36	1	1	1	0	0	0	0	1	0	0	0	0	0	1
37	1	0	0	0	0	0	0	1	0	0	0	0	0	1
38	1	0	0	0	0	0	0	1	0	0	0	0	0	0
39	1	0	0	0	0	0	0	1	0	0	0	0	0	0
40	1	0	0	0	0	0	0	1	0	0	0	0	0	0

For all the structures, topological indices (including distance, detour and Cluj indices) have been computed (Table 3) by using TOPOCLUJ software [20].

Table 3. Global descriptors and log P for the set of flavonoids in Table 1

Structure	IE max	IP max	HOMO	SD	log P
1	99	606.5	-8.584	0.708	3.4
2	119	712.5	-8.761	0.226	2.9
3	103.5	627	-8.743	-0.160	2.5
4	125	769.5	16.020	-0.257	2.4
5	179	1063	-5.395	-0.450	2.2
6	263	1524	-4.962	0.322	3
7	99	594.5	-9.527	0.611	3.3
8	188	1119	-5.426	-0.160	2.5
9	99	591	-9.464	0.612	3.3
10	107	617.5	-9.687	0.419	3.1
11	122	696.5	-7.067	-0.257	2.4
12	114	693	-6.823	0.028	2.7
13	131	766.5	-9.584	-1.151	1.5
14	134	783	-7.762	-0.257	2.4
15	129	799.5	-9.175	0.515	3.2

Structure	IE max	IP max	HOMO	SD	log P
16	192	1159	-7.423	-0.319	1.4
17	135.5	783	-7.809	-1.411	1.4
18	135.5	823.5	-6.096	-1.313	1.1
19	158.5	972	-9.275	-0.572	1.1
20	136	807	-8.829	-0.855	1.8
21	119	712.5	-8.753	0.226	2.9
22	179	1044	-8.978	0.322	3
23	149	856.5	-6.833	0.129	2.8
24	153	946	-6.756	0.129	2.8
25	200	1154.5	-7.090	1.932	4.4
26	119	726.5	-8.270	0.147	1.8
27	123.5	760.5	-9.031	-0.450	2.2
28	148.5	894	-5.375	-0.546	2.1
29	146	887	-7.361	0.603	2.3
30	179	1044	-7.644	1.287	4
31	197	1216	-7.864	0.648	3.2
32	160.5	992	-8.134	0.405	3
33	125	769.5	-8.194	0.033	2.7
34	160.5	992	-8.244	0.405	3
35	142	871	-8.075	-0.494	2
36	234.5	1447.5	-7.262	1.283	3.8
37	215	1273	-5.723	0.206	2.9
38	119	726.5	-8.302	-1.222	1.4
39	119	750	-9.419	0.033	2.7
40	146	887	-7.380	-0.643	2

3.2. Data reduction and correlation weighting

In the step of data reduction, all the descriptors with the variance Var<30% and those with intercorrelation larger than 0.80 have been discarded.Correlation weighting was performed on all the positions in the hypermolecule, nest the statistically non-significant positions were discarded. In case of the significant positions, the correlating coefficients are used to compose new local descriptors, by multiplying with the local weighted vectors (thus resulting new weighted vectors). Next, the local correlating descriptors are summed to give a global descriptor, denoted SD. This new descriptor, that is a linear combination of the local correlating descriptors for the significant positions in the hypermolecule (i.e. H13, H14, H19, H22, H27, H30 – Table 4),will be used as the basis of modeling log P (see below).

Table 4. Correlation weighting descriptors (see text)

M_i	H13	H14	H19	H22	H27	H30	SD_i
1	2.119	-1.411	0	0	0	0	0.708
2	0.675	-0.449	0	0	0	0	0.226
3	-0.479	0.319	0	0	0	0	-0.160
4	-0.769	0.512	0	0	0	0	-0.257
5	-1.346	0.896	0	0	0	0	-0.450
6	0.964	-0.642	0	0	0	0	0.322
7	1.830	-1.218	0	0	0	0	0.612
8	-0.480	0.319	0	0	0	0	-0.160
9	1.830	-1.218	0	0	0	0	0.612
10	1.253	-0.834	0	0	0	0	0.419
11	-0.769	0.512	0	0	0	0	-0.257
12	0.090	-0.065	0	0	0.003	0	0.028
13	-3.107	2.068	0	0	-0.113	0	-1.151
14	-0.769	0.512	0	0	0	0	-0.257
15	1.541	-1.026	0	0	0	0	0.515
16	-0.866	0.532	0	0.015	0	0	-0.319
17	-3.656	2.245	0	0	0	0	-1.411
18	-4.172	3.011	0	0	-0.151	0	-1.313
19	-1.557	0.956	0	0.028	0	0	-0.572
20	-2.307	1.536	0	0	-0.084	0	-0.855
21	0.675	-0.449	0	0	0	0	0.226
22	0.964	-0.642	0	0	0	0	0.322
23	0.386	-0.257	0	0	0	0	0.129
24	0.386	-0.257	0	0	0	0	0.129
25	5.006	-3.075	0	0	0	0	1.932
26	0.464	-0.309	0	-0.008	0	0	0.147
27	-1.346	0.896	0	0	0	0	-0.450
28	-1.635	1.088	0	0	0	0	-0.546
29	1.908	-1.270	0	-0.035	0	0	0.603
30	3.851	-2.564	0	0	0	0	1.287
31	1.541	-1.026	0.133	0	0	0	0.648
32	0.964	-0.642	0.083	0	0	0	0.405
33	0.098	-0.065	0	0	0	0	0.033
34	0.964	-0.642	0.083	0	0	0	0.405
35	-1.775	1.281	0	0	0	0	-0.494
36	3.274	-2.179	0.282	0	0	-0.094	1.283
37	0.675	-0.449	0	0	0	-0.019	0.206
38	-3.656	2.434	0	0	0	0	-1.222
39	0.098	-0.065	0	0	0	0	0.033
40	-1.924	1.281	0	0	0	0	-0.643

4. RESULTS AND DISCUSSION

4.1. QSAR models

The models were performed on the training set (the first 25 structures in Table 1) and the best results are listed below and in Table 5. The number of descriptors was limited to four, to fulfill the considerations of Topliss and Costello [23].

(i) Monovariate regression

$$\log P = 2.581 + 1.052 \times SD$$

(ii) Bivariate regression

$$\log P = 3.266 + 1.069 \times SD - 0.001 \times D3D$$

(iii) Three-variate regression

$$\log P = 5.099 + 0.934 \times SD - 0.003 \times Detour + 0.017 \times IE_{\max}$$

(iv) Four-variate regression

$$\log P = 4.995 + 0.932 \times SD + 0.042 \times Detour - 0.002 \times IE_{\max} - 0.005 \times IP_{\max}$$

Table 5. Best models for log P in the training set of flavonoids in Table 1.

	Descriptors	R^2	Adjust. R^2	St. Error	F
1	**SD**	**0.884**	0.879	0.279	176.25
2	**N (no. heavy atoms)**	0.041	0.001	0.804	0.991
3	**Detour**	0.034	0.008	0.807	0.809
4	**IE max**	0.001	0.043	0.821	0.018
5	**SD, D3D**	**0.914**	0.906	0.246	116.691
6	**SD, Detour**	0.906	0.897	0.258	105.648
7	**SD, Distance**	0.901	0.892	0.263	100.461
8	**SD, IP max**	0.898	0.889	0.268	96.983
9	**SD, IE max**	0.895	0.886	0.272	94.014
10	**SD, Detour, IE max**	**0.937**	0.929	0.215	105.092
11	**SD, IE max, D3D**	0.932	0.923	0.223	96.497
12	**SD, Detour, D3D**	0.920	0.909	0.242	80.897
13	**SD, IE max, IP max**	0.918	0.906	0.245	78.592
14	**SD, Distance, IE max**	0.914	0.902	0.252	74.255
15	**SD, Detour, HOMO**	0.907	0.894	0.262	68.326
16	**SD, Distance, HOMO**	0.902	0.888	0.268	64.692
17	**SD, Detour, IE max, IP max**	**0.950**	0.941	0.195	96.555
18	**SD, D3D, Detour, IE max**	0.939	0.927	0.217	77.033
19	**SD, IE max, IP max, HOMO**	0.923	0.908	0.243	60.388

One can see that, in monovariate regression, the usual alignment free descriptors, i.e. topological indices (Table 5, entries 2 to 4) correlate badly with log P; the only suitable descriptor is SD, that fits in the statistical hyperspace by an alignment procedure on the hypermolecule.

4.2. Model Validation

(a) External Validation

The values log P for the test set of flavonoids (Table 1, last 15 structures) were calculated by using the best equation in Table 5, entry 17. Data are listed in Table 6 and the monovariate correlation:

$\log P = 0.054 + 1.004 \times \log P_{calc.}$; n=15; R^2=0.768; s=0.366; F=42.956

is plotted in Figure 2.

Table 6. Calculated values of log P for the molecules in the test set(Table 1)

Molecules	log P calc.	log P
26	2.671	1.8
27	2.164	2.2
28	1.987	2.1
29	2.959	2.3
30	3.807	4
31	2.947	3.2
32	2.796	3
33	2.554	2.7
34	2.796	3
35	1.817	2
36	3.410	3.8
37	2.604	2.9
38	1.395	1.4
39	2.425	2.7
40	1.798	2

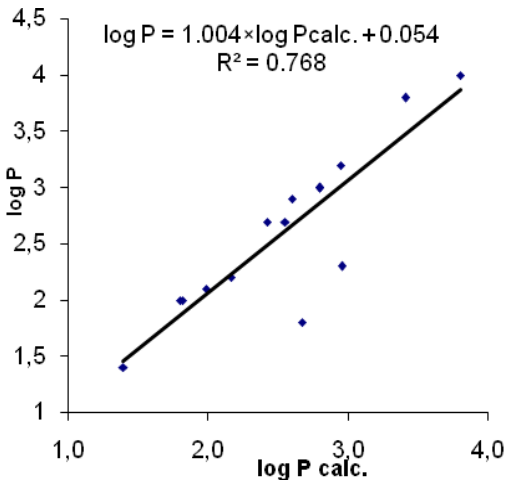

Figure 2. The plot Log P vs. log P calc. for the test set (external validation)

Similarity Cluster Validation

Validation can be performed by calculating log P for the molecules in the test set with equations learned on clusters of similarity: each of the 15 molecules is the leader in its own cluster, selected by (2D) similarity among the 25 structures of the initial learning set. The values log P calc. for each of the 15 molecules in the test set were computed by 15 new equations (the leader being left out) with the same descriptors as in eq. 17, Table5. Data are listed in Table 7 and the monovariate correlation: $\log P = -0.005 + 0.98 \times \log P_{calc.}$; n=15; R^2=0.951; s=0.168; F= 252.005 is plotted in Figure 3.

One can see that the prediction of log P by the similarity clusters is far better than that obtained in the external validation, even in the learning set (R^2=0.951 vs 0.950; s=0.168 vs 0.195 and F=252.005 vs 96.555).

Table 7. Calculated values of log P by similarity clusters, for themolecules in the test set

Molecules	log Pcalc.	log P
26	2.246	1.8
27	2.212	2.2
28	2.037	2.1
29	2.614	2.3
30	3.902	4
31	3.325	3.2
32	3.086	3
33	2.701	2.7
34	2.957	3
35	2.157	2
36	3.995	3.8
37	2.801	2.9
38	1.319	1.4
39	2.702	2.7
40	1.923	2

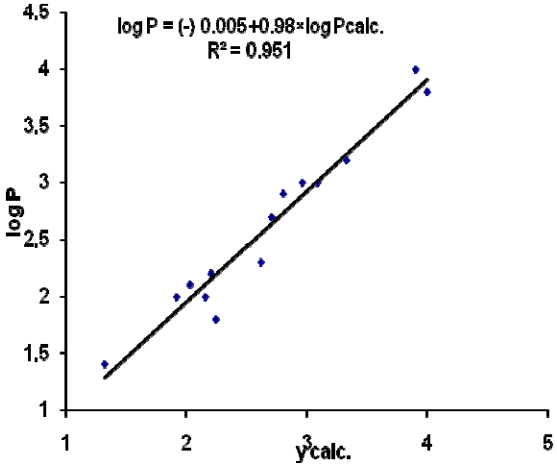

Figure 3. The plot Log P vs. log P calc. by similarity clusters

The explication of this exceptional result is that, by clustering, one obtains a set of quasi-congeners, thus making possible the basic paradigm of QSAR: similar structures show similar properties. The cluster populations can be varied to obtain the best estimation within each cluster and thus a best prediction. This represents a new correlating procedure, we call “direct prediction” and it can be performed even without previous learning steps.

CONCLUSIONS

A set of 40 flavonoids, downloaded from the PubChem database, was submitted to a qsar study by using the hypermolecule concept, in a procedure similar to that of the „alignment” of drug molecules to the biological receptors. In fact, the hypermolecule mimics the investigated correlational space, within the correlation weighting analysis. The best models have been validated in the external test set and in a new version of validation/prediction by using clusters of similarity, that favorise apparition of „quasi-congeneric” state, mandatory for a best correlation.

ACKNOWLEDGEMENTS

I am thankful fo Professor Mircea V. Diudea for helpful discussion.

REFERENCES

1. P.G. Pietta, *J Nat Prod,* **2000**, *63*, 1025.
2. K. Wolfe, X. Wu, R.H. Liu, *J Agric Food Chem,* **2003**, *51*,609.
3. A. Scalbert, G. Williamson,*J Nutr,* **2000**, *130*, 2073.
4. S.A. Payán-Gómez, N. Flores-Holguín, A. Pérez-Hernández, M. Piñón-Miramontes and D. Glossman-Mitnik, *Chemistry Central Journal*, **2010**, *4*, 12, 1.
5. C. Hansch, P.P. Maloney, T. Fujita, R.M. Muir, Nature, **1962,** *194*, 178.
6. C. Hansch, D. Hoekman, A. Leo, D. Weininger, C.D. Selassie, *Chem. Rev.,* **2002**, *102*, 783.
7. C.D. Selassie, R. Garg, S. Kapur, A. Kurup, R.P. Verma, S.B. Mekapati, C. Hansch, *Chem. Rev*. **2002**, *102*, 2585.
8. R.P. Verma, A. Kurup, S.B. Mekapati, C. Hansch, *Bioorg. Med. Chem*. **2005b** *13*, 933.
9. C. Hansch, A. Leo, *American Chemical Society: Washington, DC*, **1995**.
10. P. Willet, *J. Chem. Inf. Model.,* **2013**, *53*, 1.

11. G.B. McGaughey, R.P. Sheridan, C.I. Bayly, J.C. Culberson, C. Kreatsoulas, S. Lindsley, V. Maiorov, J.-F. Truchon, W.D. Cornell, *J. Chem. Inf. Model.,* **2007**, *47*, 1504.
12. R.P. Sheridan, *Drug Discovery*, **2007**, *2*, 423.
13. R.P. Sheridan, S.K. Kearsley, *Drug Discovery Today,* **2002**, *7*, 903.
14. P. Willett, *J. Chem. Inf. Comput. Sci.***1998**, *38*, 983.
15. A.A. Toropov, A.P. Toropova, *Internet El. J.Molec.Design*, **2002**,*1*, 108.
16. A.A. Toropov, A.P. Toropova, *J. Mol. Struct*. (*Theochem*) **2001**, *538*, 287.
17. M. Randić, *J. Chem. Inf. Comput. Sci.* **1995**, *35*, 373.
18. M. Randić, *New J. Chem*. **1995**, *196*, 781.
19. L.B. Kier, L.H. Hall, *J. Pharm. Sci*. **1981**, *70,* 583.
20. O. Ursu, M. V. Diudea, TopoCluj software program. Babes-Bolyai University, Cluj, 2005.
21. PubChem database, accessed 23.05. 2013.
22. A.T. Balaban, A. Chiriac, I. Motoc and Z. Simon, *Steric Fit in QSAR* (*Lectures Notes in Chemistry*, Vol. 15), Springer, Berlin, **1980**, Chap. 6.21.
23. J.G. Topliss and R.J. Costello, *J. Med. Chem*. **1972**, *15*, 1066.